

Facility Registry System: Data Steward Manual

May 30, 2000 - DRAFT

I. FORWARD

Data Integration and Public Access are critical activities for EPA. The ability for EPA staff and the public to have information on environmental activities at places of interest is of critical importance to being engaged in environmental protection. All this rides on the premise of high quality data in our program systems. A dedicated commitment to high data quality is displayed by having committed attention to the data. This Manual identifies the roles and responsibilities of Data Stewards to the Facility Registry System (FRS).

II. BACKGROUND

FRS Background: The purpose of the Facility Registry System (FRS) project is to provide the Environmental Protection Agency (EPA) with an authoritative central database of facility identification records that links all facility oriented program system records.

Development of the FRS supports the objective of providing environmental managers and the public with convenient access to a wide range of information necessary for effective risk-based decision making and multi-media analysis. Traditionally, EPA data systems have been developed by individual program offices to support the administration of distinct environmental programs authorized under specific environmental laws. Information is rarely transferable from one system to another without a great deal of customized intervention. This makes management of information regarding regulated facilities difficult to integrate across environmental programs.

Development of the FRS will help the Agency achieve a number of business objectives:

- Enhance internal and external access to EPA and State data and information;
- Improve data quality;
- Enhance the ability to conduct integrated and tabular data analyses;
- Facilitate error detection and correction;
- Enable direct facility review and confirmation of records;
- Facilitate back-end data administration and records management;
- Create a single point within EPA to house State Facility Master Records;
- Reduce input burden for States and regulated industry; and
- Reduce duplication and inconsistency among national data system administrative elements.

The FRS will augment several other EPA initiatives, including Central Data Exchange, Electronic Reporting, Data Standards, the Integrated Error Correction Process, and the Facility Linkage Application. Together, these activities and applications constitute the operational core of the

National Environmental Information Exchange Network (NEIEN). Activities related to FRS are described in Exhibit 1, below.

The FRS will be managed by the Office of Environmental information (OEI), Office of Information Collection.

The Data Steward Network: The history of the Facility Index System (FINDS) and other data-based initiatives has shown that an accurate, national database of facilities cannot be established and maintained by analysts at EPA Headquarters alone. The FRS system will only be used and useful if it contains accurate, high quality data. While a central facility registry must be maintained at the national level to insure nationwide coverage and universal access, those closest to the facilities contained in the database are best positioned to insure that records are complete and information is accurate. Staff at all levels with a day-to-day working knowledge of the subject facilities and data systems play a key role as stewards of the data. While automated record linkage is a necessary step, it will be individuals - Data Stewards - with working knowledge of the data and the data management systems that will underpin the effort to correct errors and maintain conditions that ensure high-quality data.

Purpose of the Data Stewards: The FRS will be supported by a network of Data Stewards at EPA and the States. The Data Stewards act as “champions” for high-quality facility data, guiding the establishment of accurate environmental linkages between facility records from different program systems based on their familiarity with places and practices in their State and/or Region.

Data stewardship is therefore an organizational prerequisite to high-quality

**Exhibit 1 –
Activities Supporting and Related to FRS**

EPA is currently developing a **Central Receiving Facility** to accept data from a variety of sources. In addition to supporting the implementation of **Electronic Reporting**, central receiving responds to industry, State, and program desires for a common point of exchange that avoids “stove-piped” data collection problems (different PINS, formats, and software). EPA is working to introduce electronic reporting for all major environmental compliance programs, both for reports submitted directly to the Agency and those submitted to State or local agencies under delegated programs.

Data Standardization will facilitate data integration needed to support multi-media environmental protection. Facility Identification is among six priority data standards. EPA has developed an interim standard for facility identification data, to be completed by the end of 2000. FRS will reflect both the Facility Identification Data Standard and the Facility Identification Template for States (FITS).

The **Integrated Error Correction Process (IECP)** will provide a standardized mechanism for submission and review of discrepancies in Agency data made available to the public. The initial roll-out of IECP will come later in 2000, and will target data made available to the public through Envirofacts. Facility-related data discrepancies found by the public will be channeled through the IECP; those associated with system data made available through Envirofacts will be “flagged” in Envirofacts.

The **Facility Linkage Application (FLA)** associates facility identification data across EPA program and State systems through computer-based name and address matching, coupled with manual reconciliation by Data Stewards. The FLA assigns an identifier that is used to associate or “link” facility data records from multiple program systems. FLA provides essential support for applications that rely on integrated views of facilities, including Envirofacts Warehouse, EnviroMapper, IDEA and OTIS.

data of any kind, including facility data. In order for individual EPA staff to be effective stewards, data stewardship must be embraced as a core value of the organizations in which those individuals are located. Since much of the data contained in the EPA data systems is compiled by the States, EPA and the States must develop an effective and robust stewardship partnership. This partnership must be premised on a mutual understanding and commitment to support each other's needs and goals. This partnership should also recognize and build upon the record of accomplishment that some States have built in the area of data stewardship and facility registration, especially through the One-Stop Program.

Data Stewards, and the process of data stewardship, will be guided by the following principles.

1. Data stewardship is a collaboration of peers based on a shared, mutual goal of achieving and maintaining high quality environmental data to promote the protection of public health and the environment.
2. Data stewardship is a collaboration among the various levels of government involved in environmental program implementation.
3. Data stewardship is based on respectful use of the data collected and shared among environmental regulatory entities.
4. Data stewardship recognizes that non-confidential and non-enforcement-sensitive information should be delivered to the public through clear, efficient, and accessible mechanisms.
5. Data should be managed to be shared across many enterprise applications; data stewards facilitate this process.
6. Data stewards should view themselves as information quality "champions," and motivate their organizations to recognize that facility data is "mission critical" information for environmental agencies.
7. Data stewards should adhere to a standard of continuous improvement in the integration of data about places of environmental interest, including all associated activities of concern and responsible parties.
8. Data stewards should leverage State information managers' localized knowledge of geographic locations, facility business, environmental regulations, and economic sectors without placing an undue burden on anyone.
9. Data stewards should identify and propose corrective actions for root problems in EPA information collection approaches and programmatic management practices that contribute to the Agency's inability to accurately associate facility data across program information systems.

III. ROLES AND RESPONSIBILITIES OF DATA STEWARDS

The FRS Data Steward Network is based upon a model of collaborative interaction among EPA Headquarters, Regional, and Program staff. The collaborative model is designed around the fact that Program activities are conducted at Headquarter and Regional levels. As mentioned previously, the Data Steward Network is also based upon the understanding and conviction that collaboration between EPA and States is key to achieving FRS and other data quality objectives.

This section describes the roles of Data Stewards in EPA's Office of Environmental Information, Media and Enforcement Programs, and Regions. More particularly, the section describes specific roles and responsibilities for a (i) Data Stewardship Program Manager, (ii) Program Data Stewardship Managers, (iii) Regional Level Data Stewards, and (iv) State Data Stewards. Each Data Steward role will be defined through its own tasks as well as its relationship with other aspects of the Data Steward Network.

Data Stewardship Program Manager: Housed in the Office of Environmental Information, the Data Stewardship Program Manager will coordinate the Data Steward Network. The Program Manager will be responsible for: (1) overseeing EPA Regional data stewardship activities, and (2) coordinating with EPA Program Data Stewards. The Program Manager:

- a.* Facilitates the resolution of issues or concerns about facility identification and linkages that arise between and/or across Regions and States.
- b.* Resolves conflicts and acts as liaison between Regional Coordinators and EPA Program Stewards to ensure timely completion of program system updates and data corrections.
- c.* Works with *Regional and Program Data Stewards* to develop customer service standards for the Data Steward Network for responding to facility linkage Discrepancy Reports. Manages the distribution of facility linkage Discrepancy Reports to Regional Data Stewards and monitors and facilitates compliance with customer service standards in responding to those reports.
- d.* Manages the distribution of Discrepancy Reports detailing inaccurate facility identification and other data concerns to EPA *Program Data Stewards* and monitors and facilitates compliance with these reports.
- e.* Works with the *Regional and Program Data Stewards* to develop a performance measurement program with appropriate statistics for the Data Steward Network. Gathers performance measurement data from Programs and Regions on regular intervals and assesses and reports Network performance.

Program Data Stewardship Managers: Housed in National Program Manager's Offices, Program Data Stewardship Managers will be responsible for coordinating with the Data Stewardship Program Manager and Regional Stewards to address data quality issues in their

respective systems. Program Data Stewards may be media programmatic in nature or systems-oriented, depending on the program. Program Data Stewardship Managers shall:

- a. Work with Regional and/or State program staff to fix erroneous and/or incomplete data in situations where the Region or State maintains the primary data source.
- b. Work with the media program managers, identify what data are collected, the data flow, and data system business rules. Identify potential changes to what data is collected and/or method by which data is reported so as to improve program management and reporting.
- c. Identify systemic problems/inconsistencies in the underlying information collections that yield data quality errors, and work with program management to find and implement corrective solutions.
- d. Respond to Discrepancy Reports distributed by the Data Stewardship Program Manager.
- e. Coordinate with *Regional Data Stewards*, gather and report to the *Data Steward Program Manager* on a regular intervals performance measurement data for the program data system.

Regional Data Stewardship Coordinators: [Note: This document outlines two distinct data stewardship roles at the Regional Level. It is recognized that some Regions may decide to combine these roles through a single individual.] Regional Data Stewardship Coordinators will have the overall responsibility for the quality of facility linkages for records located within their respective Regions. They are also the point-of-contact for other *Regional Data Stewardship Coordinators*, internal media specific program managers and *Program Data Stewards*, and the *Data Stewardship Program Manager*. Regional Data Stewardship Coordinators will also serve as primary point-of-contact for State-level stewardship activities. Specifically, the Regional Coordinators will:

- a. Participate in State/EPA work groups to develop strategies for how liaisons with States should be established within their Region. Serve as a point-of-contact for the Region with State and EPA Headquarters data stewards. This will include communicating the implementation of the FRS in the Region; coordinating FRS, Central Data exchange, and other related initiatives.
- b. Approve access by Regional and State personnel to the Facility Registry System. Coordinate necessary training activities for Regional and State users.
- c. Coordinate with other Regional Coordinators to develop strategic plans for data linking, reconciliation, quality, and clean-up. Coordinate Regional priorities with State priorities for data clean-up where appropriate.
- d. Notify Program Data Stewards of data quality issues that are raised by Regional or State

data stewards.

- e. Report to the EPA Data Stewardship Program Manager needed alterations or enhancements to the FRS.
- f. Coordinate with relevant Program staff working within the Regional Office and/or the *Program Data Steward*, conduct ongoing audits to evaluate stewardship performance and communicate with the *Data Stewardship Program Manager* at regular intervals.

Regional Data Stewards: [Note: This document outlines two distinct data stewardship roles at the Regional level. It is recognized that some Regions may find it necessary or desirable to combine these roles through a single individual.] Regional Data Stewards will support the *Regional Data Steward Coordinator* in the Region and are a point-of-contact for *Program Data Stewards* and *State Data Stewards*. The *Regional Data Stewards* shall be responsible for the following:

- c. Maintain contact with all *State Data Stewards* to ensure agreed to State-EPA interactions occur as expected.
- d. Communicate new/updated valid values with the Data Stewardship Program Managers for placement in the FRS.
- e. Perform such activities as are necessary to ensure accurate linkages of data for a facility across program systems. The activities will include:
 - S Where one facility identification number has been assigned to more than one facility, the data stewards must break the incorrect linkages using *Move* procedures to ensure that a transaction history is maintained.
 - S Where more than one facility identifier has been assigned to one facility, the data stewards must link the facility records appropriately, using *Merge* procedures that ensure that a history is maintained of the transaction.
 - S Review Discrepancy Reports and make corrections to facility linkages as necessary; coordinate with appropriate Program staff to acknowledge receipt and take necessary actions.
 - S Process candidate linkage files and make the appropriate changes. Prepare summary reports generated from the FLA explaining all changes to linkages.
- d. Compile and report the performance measurement data through the *Regional Data Steward Coordinator*.

Participating State Data Stewards: State Data Stewards will have analogous responsibilities as the EPA regions. How a State agency decides to distinguish these roles or what titles they use

shall be determined by the state. Roles and responsibilities can include:

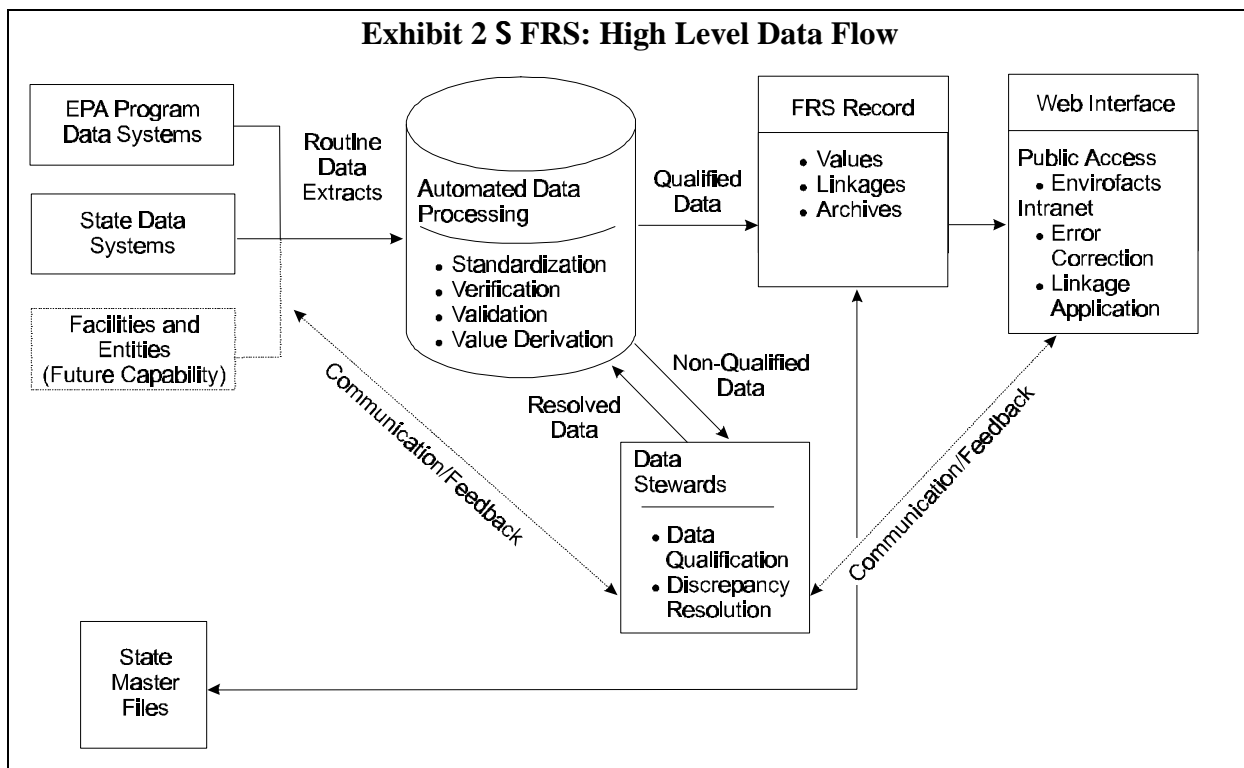
- a.* Participating in State/EPA workgroups and serving as points of contact for their State.
- b.* Establishing mutually agreeable relationships with Regional coordinators/stewards.
- c.* Establishing internal networks that access and leverage the knowledge of staff with multi-media facility knowledge for participation in the facility resolution process. What this network looks like will be dependent on the individual state agency's business needs and organizational culture.
- d.* Performing the manual processing and resolution of candidate linkage files for their State facility records.
- e.* Working with the Regional coordinators to address state priorities as well as EPA Regional and Headquarters priorities.

IV. PROCEDURES FOR REGIONAL DATA STEWARDS AND DATA STEWARDSHIP COORDINATORS

Data Steward activities will tend to fall into four basic categories: (1) investigating and resolving issues with data records that are being or have been processed by the system; (2) resolving issues with completed records; (3) communication and coordination with other Data Stewards to maintain and improve the effectiveness of the Data Steward Network; and (4) FRS system administration and training activities.

The section is organized around the four basic categories of Data Steward activity. Part 1, "Investigation of Data Records and Resolution of Issues," reviews major automated system functions, and describes procedures that Data Stewards should employ to investigate and resolve discrepancies and/or incomplete records. Part 2, "Resolving Issues with Completed Records," describes the business rules that determine the form and content of data elements in the FRS record. Part 3, "System Administration," describes Data Steward responsibilities regarding system access, training, and other user needs. Part 4, "Communication and Coordination Protocols," describes recommended practices to assure adequate communication among various levels of data stewardship. This section also describes the Integrated Error Correction Process (IECP), a web-based application that will enable Data Stewards to submit information to the FRS system.

A generalized depiction of the FRS process is provided in Exhibit 2. Exhibit 3 is a side-by-side schematic, showing how data steward activities correspond to key nodes and functions of the FRS system.



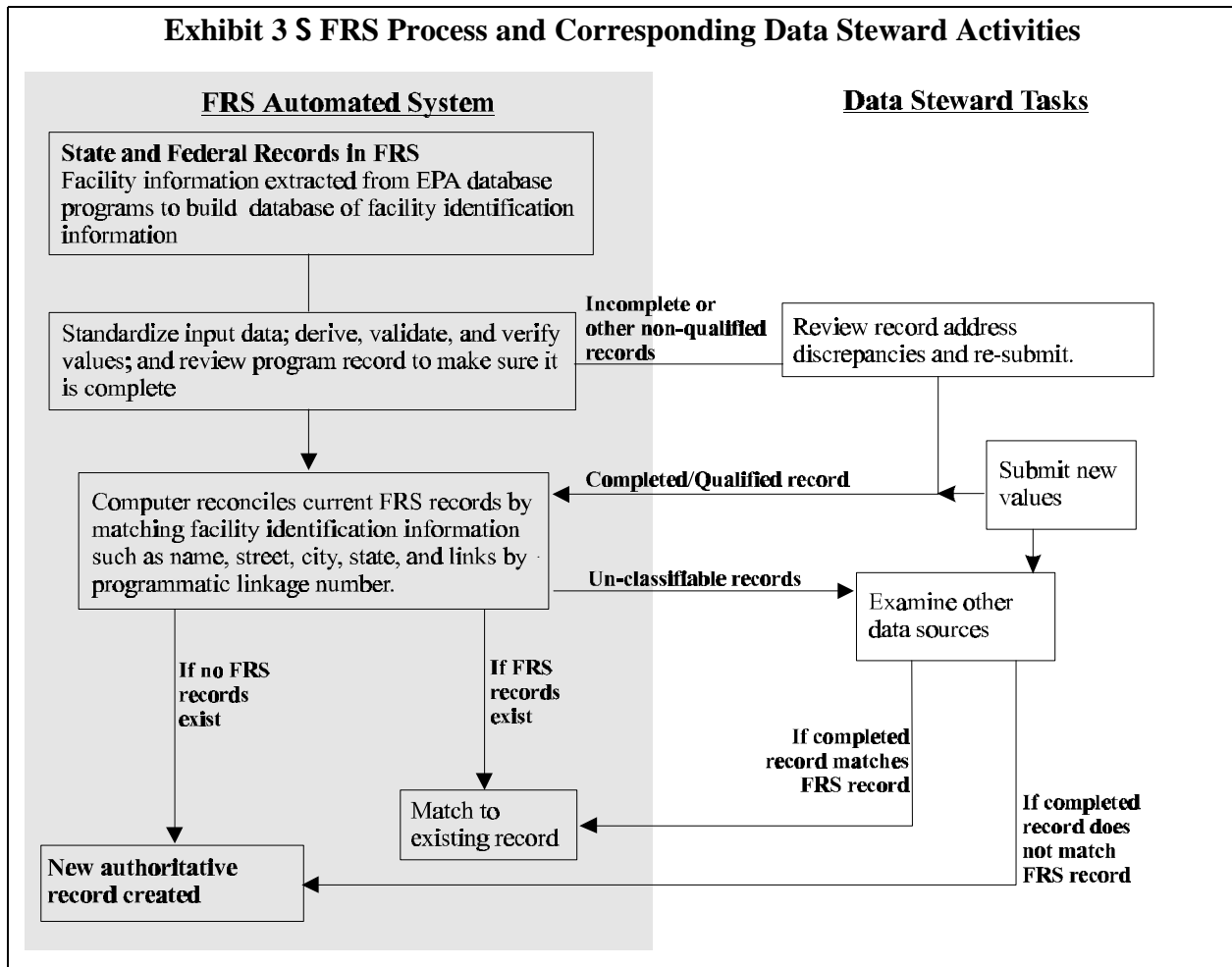
Appendix 1 contains a list of definitions pertinent to Data Steward activities and the FRS process; Appendix 2 contains a detailed table of all 63 FRS data elements and associated business rules; and Appendix 3 contains a list of references relevant to data stewardship and the FRS process.

Part 1 – INVESTIGATION OF DATA RECORDS AND RESOLUTION OF ISSUES

The FRS system will extract facility data from State Master Records and EPA Program systems. (See FRS Web Site at: WWW.EPA.GOV/ENVIRO/XXXX for a set of data mapping models and documents which explain the mapping from the national systems to the FRS record.) Facility site name and address data elements are standardized to aid in automated matching processes. Data quality is checked using standard reference tables. As indicated in Exhibit 3, incomplete facility files that do not pass validation checks will be flagged and must be reviewed by the Data Steward manually.

Data Source Prioritization S When attempting to complete missing data or reconcile conflicting data, Data Stewards should adhere to the following data source prioritization, with sources listed first assuming a higher, or more authoritative, stature; State Master Files, Central Receiving Certificate, TRIS, RMP, RCRA TSD, PCS Major, AFS Major, RCRA LQG, SF-NPL, PCS minor, AFS non-Major, RCRA SQG, PADS, Superfund-CERCLIS, Docket, SSTs, and NCDB. When determining data source prioritization, it is also essential to consider the time-relevance and/or vintage of data from a particular system; with newer data being viewed as more authoritative than older data.

Exhibit 3 S FRS Process and Corresponding Data Steward Activities



This section is broken into three sub-parts: (A) Incomplete Records; (B) Duplicate Checking; and (C) Data Quality Checking. Each of these sub-parts is further divided into two sections: automated activities and manual activities. The automated activities briefly describe the process that the FRS system will use to organize and process the data. The manual activities describe the process the Data Steward should use in completing or organizing data flagged by the FRS.

A. Incomplete Records

Automated Checks: The FRS checks data for completeness before loading. A complete FRS record holds 64 distinct data elements. A record is considered incomplete if any of these elements are missing or some “equivalent of missing.” “Equivalent of missing” is defined as the existence of certain values or anomalies (e.g., “UNKNOWN,” “N/A,” “NAME NOT KNOWN”) that are considered to be no better than if the values were missing. A look-up table contains a list of invalid name patterns and the contents of the data elements are verified against this table for completeness. The authoritativeness and overall quality of the FRS record is ensured (in part)

through a process of automated cross-checks to verify critical relationships among key data elements. Although FRS will load incomplete data, it will not be considered as “verified.” In other words, the more complete the FRS record, the more confident one can be regarding the accuracy of any particular aspect of the record.

It is especially important that data values exist, or are derived, for the following attributes:

- Facility Site Name, Location Address, Locality Name, County Name, State Name, Country Name (if outside the United States), and Location ZIP Code. Note: If the location address contains a mailing address (“P.O. Box” reference), it is considered incomplete.

FRS will produce a report, for manual review, which indicates which facilities did not have complete data. It is the Data Steward’s responsibility to obtain or derive missing data values, complete the record, and re-submit the new data package through the Integrated Error Correction Process.

Manual Resolution: The Data Steward will evaluate the report indicating which facilities do not have complete data and attempt, working with other data stewards in the Programs, Regions, and States, to gather the data necessary to complete the record. Compiled in Table 1, potential sources of missing data include Dun and Bradstreet data and various Internet resources, such as, the Web White Pages, Web Yellow Pages, the U.S. Post Office Web site, Map Quest web site, SEC-EDGAR web site, the Department of Energy Web site, various DOD web sites, and other credible resources.

In all cases, it is important to work with the appropriate State and/or Program Data Stewards, both as a potential source of missing data and to assure they are fully and promptly apprized of FRS activities. Once the missing data has been obtained, submit the revised record through the Integrated Error Correction Process.

Table 1 - Completing An FRS Record: Some Key Web Resources

Resource Organization	URL
Dun & Bradstreet	www.dnb.com
Web White Pages	www.whitepages.com
Web Yellow Pages	www.YellowOnline.com or www.WorldPages.com
U.S. Postal Service	http://new.usps.com
Map Quest	http://www.mapquest.com
Securities & Exchange Commission	www.sec.gov/edgarhp.htm
U.S. Department of Energy	www.doe.gov
EPA Envirofacts	www.epa.gov/
USGS: Geographic Names Information System	http://mapping.usgs.gov/www/gris/

B. Duplicate Checking

Automated Checks: To ensure that one facility site is not assigned multiple facility identification numbers, checks are used to identify facility sites where assignment of a unique identifier could be considered questionable. Prior to adding a new facility site (the facility site does not already exist in FRS), the FRS system will:

- Potential Matches: Search the database (both facility sites and environmental interests), using the Program Information System Identification Number and name and address matching algorithms, to identify “matches” and “potential matches”. If a “match” is found, FRS will create an Environmental Interest and pair it to the matched FRS Facility Site. If a “potential match” is found, FRS will create a report that identifies the facility data and its potential matches for manual review. Data Stewards will review these reports and conduct research to either confirm or “disqualify” these potential matches.
- Multiple Facility Records in One or More Data Systems: Identify facility sites where more than one facility site has reported the same Environmental Interest identifier. For example, if two different Toxic Release Inventory System (TRIS) facility sites report the same Resource Conservation and Recovery Act Information System (RCRIS) identifier, it may be postulated that the facility site is being treated as two sites by the TRIS program and as one site by the RCRIS program. This case must be resolved manually.

Manual Resolution: Data Stewards must promptly review FRS system reports on potential matches to either confirm the relationship or identify duplicates. Data Stewards supporting the Envirofacts Facility Linkage Application (FLA) have identified duplicate entries through manual cross-checks of at existing facility identification data (i.e., alternative names, supplemental address, mailing address, and DUNS number). This process is performed for each “potential match”. Based on the similarities, differences, and other pertinent data, a judgement is made, and the resulting facilities are linked together in the FLA application. Data Stewards are encouraged to follow a similar approach in support of the FRS. When the Data Steward has confirmed the match or identified a duplicate record, use the Integrated Error Correction Process to submit your record revision to the FRS.

As part of this review, the Data Steward is encouraged to access and review the environmental and geographic data available through Envirofacts. For example, if a TRIS facility site reports releases of a certain chemical to air, a linkage to an Aerometric Information Retrieval System/AIRS Facility Subsystem (AIRS/AFS) environmental interest should be established. If a potential match is an AIRS/AFS facility site and the environmental data for the AIRS/AFS facility supports releases of the same chemical, additional weight is added in support of matching the two sites together.

C. Data Quality Checking

Automated Checks: FRS will perform automated quality checks on all entered data. The FRS will identify and correct invalid data and will produce a report that identifies all inconsistencies found that could not be corrected during the automated review process. A repository of look-up information is used to validate data. For a more detailed discussion of the sources of information used to validate data, see WWW.EPA.GOV/ENVIRO/FRS.

The data quality checks to be performed on specific data elements are described by entity below.

Facility Site: The physical location attributes (i.e., Country Name, State Name, County Name, Locality Name, and ZIP Code) are validated using FIPS 55-DC3 and the USPS ZIP Code reference tables. The FIPS file was downloaded from the official FIPS web site, which was last updated in 1998. This file provides a two-character state code and five-character numeric place code to uniquely identify each listed entity. An exhaustive list is carried of incorporated places, census designated places, primary county divisions, recognized Indian reservations and Alaska Native villages, and counties. The USPS ZIP Code file is mainly used to verify postal codes in county boundaries. Each data element is checked both individually (e.g., does the county exist in the state?) and in combination across data elements (e.g., is the locality/county combination valid within the state?).

Geographic Coordinates: All data quality checks on latitude/longitude data will occur within the Locational Reference Tables (LRT) of Envirofacts. Latitude/longitude data is mapped to and subsequently processed and stored within the LRT, which serves as a central repository of latitude/longitude data. The LRT uses computer algorithms to check the coordinate data for compliance with Agency locational data standards. Typically,

there may be several different sets of coordinates for a single facility site. LRT includes a function that determines the best set of coordinates to use for mapping a facility site, which is based on the accuracy value of each coordinate, and a ZIP and county boundary check. The recommended coordinates are marked by a “Best Value Flag”. The FRS facility site will be linked to the set of coordinates marked by the “Best Value Flag”.

Environmental Interest: The Environmental Interest Type Name will be validated using a set of permitted values maintained in a reference table. The Environmental Interest Type indicates the environmental permit or regulatory program that applies to the facility (e.g., TRI Report, National Pollutant Discharge Elimination System [NPDES] Major, Risk Management Plan [RMP] Facility). The Information System Abbreviated Name is validated using a set of permitted values maintained in a reference table. The Information System Abbreviated Name represents the name of an information management system for an environmental program.

Self-reported linkages (Information System Identification Numbers) will be verified as existing in Envirofacts/FLA. An additional check determines if facilities are linked in FLA. If they are, then these two facilities were linked together either through name and address matching or by the Data Steward, which provides sufficient verification. If they are not linked in FLA, they will be flagged for manual review.

SIC and NAICS Codes: SIC codes are validated as existing in the standard SIC reference. NAICS codes are validated as existing in the standard NAICS reference.

Manual Resolution: As described above, the FRS will flag inconsistencies identified through the automated review for resolution by Data Stewards. Data Stewards shall review the FRS data quality discrepancy report, verify the accuracy of data values, and submit findings through the Integrated Error Correction Process. As described under Part 1, sub-heading “A,” above, the Data Steward should utilize available resources such as those identified in Table 1 to confirm data values. When necessary or appropriate, the Data Steward should confirm findings through timely communication with relevant State/Program Data Stewards.

Part 2 – RESOLVING ISSUES WITH COMPLETED FRS RECORDS

The FRS will be populated with regular data extracts from Program Systems, State Master Files, Central Data Exchange certificates, and other respected sources such as Dun and Bradstreet, U.S. Postal Service Zip Code reference tables, USGS State and County code reference tables, and internet sources such as the Web White or Yellow Pages. To the extent possible, the FRS will reflect the most current, accurate, and up-to-date information about the place of interest. To ensure the integrity of these updates, a set of business rules have been developed to guide how and when FRS records can be changed. Some changes will be accomplished through the automated resolution process, other changes will require attention and manual intervention from Data Stewards.

FRS data elements are defined in Appendix2. The following specifies the update business rules applicable to FRS data elements when manual intervention will be required or expected.

Data Element: **Facility Site Name**

EPA receives records with varying facility site names from a variety of sources. Facility Site Names could change as each subsequent records are used to enhance the FRS record and update the linkage. However, all names from the source systems will be archived in the FRS and will be searchable.

The FRS records will be populated with the facility site name from the first system used to populate the FRS. Data Stewards will be asked to manually intervene if the facility site name is substantively different on subsequent submissions from program system refreshes. Manual research might include using internet resources such as the Web White or Yellow pages, accessing corporate web sites, communicating with other Regional or State Data Stewards, or checking name information available through Dun and Bradstreet. The FRS record would be modified upon completion of the manual research by the Data Steward. Updates from State Master Files and Central Data Exchange certificates would normally not trigger manual intervention by the Data Steward even if the facility site name were substantially different if the currency of the update was more contemporary than the record in the FRS and the automated process could assure that the new name was the same place as the old name. State Master Files and Central Data Exchange Certificates are expected to carry high levels of integrity. As new records are received, the FRS application will update the facility site name data element source information and the data element would be time-stamped.

Data Elements: **Location Address, Supplemental Location Text, Locality (or City) Name, State, Zip Code, County Name, Country Name.**

Data Stewards will be asked to conduct research on location address data if the FRS receives a new record that is a very strong “potential match” with an existing FRS record, i.e., same facility site name, same contact names, same mailing address, same affiliation information, but the location address data is incomplete or partially invalid. Manual research might include looking at a map of the area, accessing the organization’s web site, or checking the organization’s address through internet resources such as the Web White or Yellow pages.

Data Element: **EPA Region Code**

Occasionally EPA Regions are assigned responsibility for places outside their geographic limits. Data Stewards might be asked to provide help in identifying these anomalies.

Data Elements: **Environmental Information System Abbreviated Name and Environmental Information System Identification Number**

Because of the Regional Data Stewards’ relationship with State program managers and State Data Stewards, they may be asked to help identify the source and appropriate data values for

identification numbers and system names that are not owned and managed by EPA's national programs.

Data Elements: **Environment Interest Start Date** and **Start Date Qualifier**, **Environmental Interest End Date** and **End Date Qualifier**

Data Stewards can provide input to the FRS program managers if the data in the FRS does not accurately reflect the interest status based on their knowledge of the program.

Data Elements associated with Affiliation information including: **Organizational Formal Name**, **Affiliation Type**, **Parent Company Name**, **DUNS Company Numbers**, **Mailing Address**, **Supplemental Address Text**, **Mailing Address City Name**, **Mailing Address State Name**, **Mailing Address Country Name**, **Mailing Address ZIP Code**, and Facility Contact Information including **Individual Full Name**, **Individual Title Text**, **Electronic Mail Address**, and **Telephone Number**.

Data Stewards are encouraged to provide updated affiliation information as they become aware of changes. If changes are submitted, the FRS will be changed and the FRS will note the Steward as the source.

Part 3 - SYSTEM ADMINISTRATION

FRS Regional Data Stewards are responsible for a range of System Administration activities, including user registration and training. Regional Data Stewards approve access by Regional and State designated personnel to the FRS/FLA.

The FRS/FLA O&M application provides an interface for new user registration and user activation and maintenance to handle user administration. Access to FLA should be restricted to registered users only, and a valid ID and password are necessary to logon to the application. Any user can request access to the O&M application through the Regional Data Steward Coordinator. The prospective user must provide the required information and stipulate the appropriate (i.e., Region, State, Program) access level. The Data Steward should review the user registration information for completeness. If the registration is approved, the Data Steward should notify the prospective user when the user ID and password have been activated.

Other system user administration activities include the following:

- The system provides an interface for authorized Data Stewards to enable or disable users, change user access, and delete obsolete users.
- The FLA application allows Regional Data Stewards to limit data administration or data access on a Program, Regional, or State basis.
- The FRS/FLA provides the capability to assign users to categories that define access levels (e.g., enable data entry for TRI users). The following information must be obtained from

the potential user to complete the registration process: name, title, agency name, mailing address, mail stop, city, state, ZIP code, office phone number, fax number, E-mail address, user ID, password, and user request category (i.e., Headquarters, Region, State).

Data Stewards are also expected to provide user support and training for the FRS and FLA applications. In particular, Data Stewards shall introduce new users to the basic system functions, provide an overview of the FRS data elements and business rules, and background on essential aspects of the FRS/FLA technical environment and system description (e.g., equipment, support software).

Part 5 - INTEGRATED ERROR CORRECTION PROCESS

The Environmental Protection Agency has developed the Correction Process (IECP). Operating through a Web interface, the IECP helps stakeholders and the public to route data errors and discrepancies to the appropriate data system for resolution. While the decision to change data remains with State or Programmatic data owners, the IECP provides a uniform mechanism and procedures for accepting input, routing and tracking discrepancies. The IECP also produces periodic management status reports.

The IECP will deal with all EPA data, making its purview broader than that of the 64 FRS data elements. However, the IECP can be used as a tool for FRS data stewardship. Appendix 4 contains draft Standard Operating Procedures for the IECP.

Completed FRS files will be used by EPA and State data systems and will be available to the public. Occasionally, a user may identify a data issue in the FRS record. For example, the user may feel that information in one of the data elements is incorrect, or that facilities with no relation are inappropriately linked. Use of the IECP process will forward the notification to the appropriate Data Steward according to the following process:

A. Intake

The issue may be brought to the attention of EPA or a state through a variety of channels, such as the IECP Web interface, telephone hotlines, e-mail, or written comments. Whatever the source, all Data Stewards should be prepared to perform an intake on a data issue. During the intake process, the Steward should attempt to gather as much of the following information as possible:

- Facility identifying information, including FRS identifying number, name, location, or other appropriate data elements, as identified in Appendix 2 – Detailed Table of FRS Data Elements;

- Problem with the record as described by the user;
- Any proposed solution as described by the user; and
- User contact information (name, e-mail, phone number, fax number).

B. Initial Investigation

The Data Steward should examine the FRS record to confirm the user's issue. Resolve the issue, if possible, through educating the user as to the meanings of various data fields or data in data fields if user misunderstanding is the source of the issue.

C. Routing

If the issue cannot be resolved through investigation of the FRS record and education, the Data Steward should use the IECP to route the issue to the appropriate Program, State, or Regional Office. The intake Steward will remain the point of contact with the user and will communicate any routing steps to the user. The IECP Web interface can be used to help the Data Steward performing the intake and the Data Steward resolving the issue to remain closely coordinated.

D. Resolution

Once the issue has been routed to the appropriate Data Steward, that Steward should investigate the record in the original program data system or systems. The Steward should investigate the issue using resources at their disposal and garnering any practical knowledge of the facility or facilities necessary that are not at the disposal of the Steward.

E. Communication

Communication with users regarding the investigation and resolution is key to effective issue management. The intake Steward should assure that all IECP communication procedures are followed and that the matter is resolved as quickly as possible. The intake Steward also should communicate any resolution, the reasoning behind that resolution, or plan to resolve the issue to the user as soon as practicable after the resolution or plan has been agreed to by the Data Stewards.

F. Tracking and Reporting

The tracking and reporting of the activities above, along with those described in Part 1 of this section are described in Appendix 4.

APPENDIX 1: DEFINITIONS

Automated Resolution: Automated resolution refers to the reliance on automated computer processes and algorithms to match like records and to fill incomplete fields based on established rules.

Data Stewards: individuals responsible for establishing and managing a process to facilitate the creation and maintenance of accurate facility information records within their sphere of responsibility and for enabling correct integration of facility information records/linkages. They are champions of information quality and successfully motivate their organizations, management, and partners to recognize the importance of improving facility identification information to meet business needs. They are chosen on the basis of their knowledge of local geography, data collected about facilities by the program systems, and economic activities of facilities. Data stewards must work together to establish and implement policies and procedures for linking facilities. Data Stewards exist at the following jurisdictional levels:

- EPA Data Stewardship Program Manager (a.k.a. Program Manager) is the central leader, overseeing and coordinating data steward activities and ensuring overall data quality for facility identification across all EPA Regions, Program Offices, and participating States.
- Regional Data Stewardship Coordinator (a.k.a. Region Coordinator) coordinates all data stewardship responsibilities and ensures overall facility data quality within a Region.
- Regional Data Stewards are assigned by the EPA to ensure data quality for facility identification throughout their own Regions, and work with State partners in accordance with procedures agreed upon by the Regional Coordinator and the primary State Data Steward.
- EPA Program Data Stewards are assigned by EPA national program managers to maintain quality data within their own program systems by responding to data quality issues identified by data stewards. These program data stewards may be located at Headquarters or the Region, may be programmatic or systems staff, depending on the nature of the program. In addition, there are other programs with cross-media responsibilities, and usually cross-media knowledge of facilities, that are encouraged to participate as Program Data Stewards.
- State Data Stewards - States manage the integration of State information collections for their own business reasons, and regulate many more facilities than those covered under Federal regulation. Many have developed their own management practices, procedures, and support systems for facility linkages. States who volunteer to participate in the Facility Data Stewardship Partnership will designate a primary Data Steward to work with

EPA in the linkage verification process for Federal and State facility records on facilities located within their State.

Derived Data: These are facility related data that the Agency collects from sources other than the facilities themselves.

Facility Identification: is the process of identifying entities responsible for or associated with activities of environmental interest that occur at a specific place.

Facility Registry System (FRS) is a central database of facility identification records that links all facility oriented program system records. The FRS includes a linkage application which links together facility identification data across EPA program and state systems through computer-based name and address matching and data steward clean-up efforts.

Making/Breaking Links (a.k.a. Linking and De-linking): Refers to a process in which relationships are established between facility identification data across EPA program and state systems through computer-based name and address matching (automated resolution) and manual review and reconciliation. A “linkage” is an identifier, assigned by an individual media program system, that generally ties a permitted activity or environmental interest to the facility. De-linking is required in those cases where incorrect matches are made by algorithms used to establish links.

Manual Reviews: Manual reviews refer to reviews performed by data stewards to assess data quality, to research and obtain missing data, to establish personal contact with sources.

Unique Identification Number (UIN): An unintelligent identification number assigned by the EPA Facility Registry System to uniquely identify a facility site.

Verification and Validation Procedures for FRS: Procedures that will be applied to facility records to ensure data quality, completeness and consistency.

APPENDIX 2 – FRS DATA ELEMENTS AND ASSOCIATED BUSINESS RULES

Recommend that this appendix list the Facility ID Data Standard data elements and definitions as well as the Facility ID Data Model. Or FRS data elements (larger than the standard) - and the FRS data model.

APPENDIX 3: REFERENCES

Facility Registry System (FRS) Validation and Verification Procedures: Work-In-Progress. Report prepared for the U.S. EPA Office of Environmental Information. SDC-0002-013-HB-2014. March 3, 2000.

Central Receiving. Briefing presented the FRS Team by Matt Leopard, OEI Collection Services Division, Central Receiving Branch. January 5, 2000.

Facility Identification Briefing. December 29, 1999.

Facility Registry System Beta Version Data Quality Report, SDC-0002-013-HB-2008. December 19, 1999.

A Data Steward's Guide to the Facility Linkage Application. Report produced by EPA Region IX. November 1999.

Facility Identification Information Business Rules: Introduction and Options Papers. Draft papers developed for the ad hoc Facility Identification Standard/Business Rules Staff Group. (Version 9/30/99).

Software Design Document for the Facility Registry System (FRS), SDC-0002-013-MM-1040, August 13, 1999. [http://oasint/rtpnc.epa.gov/frs/FRSDESIGN\\$.Startup](http://oasint/rtpnc.epa.gov/frs/FRSDESIGN$.Startup)

Proposed Facility Identification Data Standard Final Draft (Version 7/21/99). <http://www.sso.org/ecos/FII%20Std7-19.htm>

Software Requirements Document for the Facility Registry System (FRS), SDC-0002-013-HB-1032. July 17, 1999.

Facility Identification: A Proposed Strategy for EPA to Work with Its State Partners. Working Draft. May 7, 1999.

Facility Registry System (FRS). Briefing presented by the Chief of the Targeting and Evaluation Branch, Office of Compliance. Discussion Meeting with Washington State. March 15, 1999.

Data Steward Roles and Responsibilities. Table developed by EPA Region III. February 26, 1999.

FY 1999 REI Integrated Program Management Plan, Task Area D: Facility Identification Initiative Phase III. <http://www.epa.gov/rei/intgtd/IPQ3/taskd-q3.htm>.

Guidance for Facility Data Stewardship Program - EPA/State Partnership. Paper produced for EPA Region V. November 24, 1998.

FII Application Description and Business Rules. Draft Paper. October 12, 1998.

Facility Identification Template for States (FITS): Working Guidelines for Integrating Facility Identification Information. Report sponsored by the Environmental Council of States, U.S. EPA One Stop Program and the Washington Department of Ecology. December 15, 1997.

Summary Report of FINDS Region 9 Data Cleanup Activities. Report prepared for the U.S. EPA Office of Information Resources Management. SDC-0055-040-MC-6025. July 18, 1997.

Integration of EPA's Facility Data for the Key Identifiers Initiative: Value of Facility Data Linkages. Draft Report prepared for the U.S. EPA Office of Information Resources Management. SDC-0051-051-LF-5042. April 26, 1996

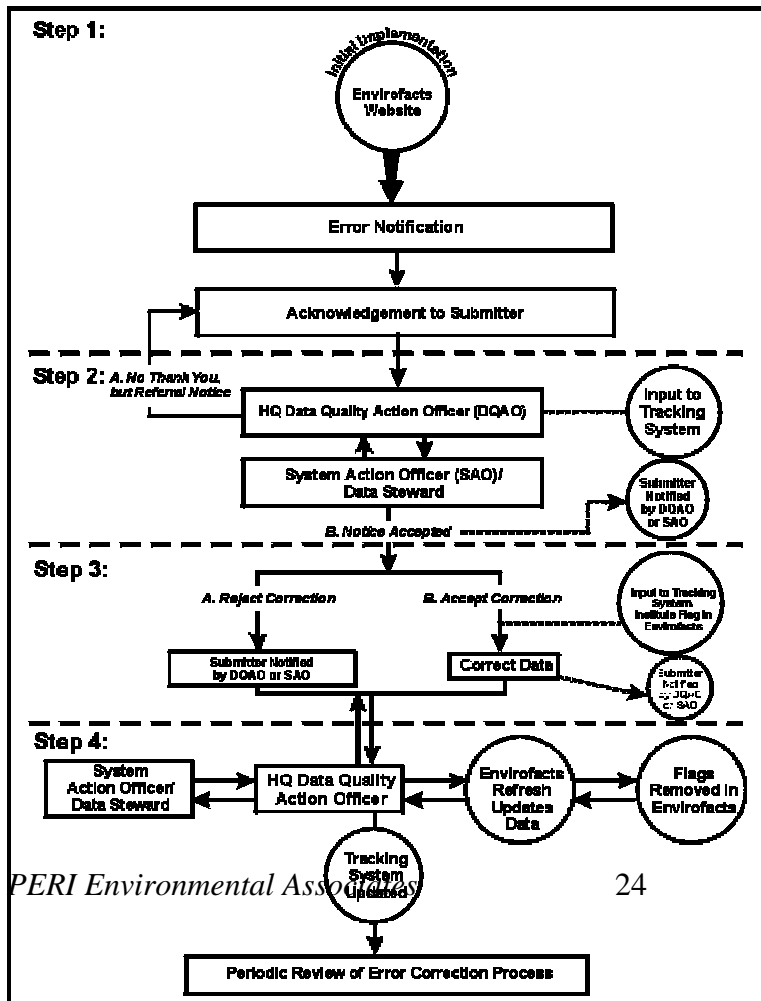
FRS Frequently Asked Questions (Draft). [http://oasint.rtpnc.epa.gov/frs/FRS_FAQ\\$.StartUp](http://oasint.rtpnc.epa.gov/frs/FRS_FAQ$.StartUp)

APPENDIX 4: Integrated Error Correction Process

Draft Standard Operating Procedures for EPA's Integrated Error Correction Process

The Environmental Protection Agency (EPA) increasingly serves the public's right-to-know about environmental quality and trends by providing access to the Agency's data. In response to internal and external concerns about the quality of EPA data, the Deputy Administrator called for development of a Data Quality Strategic Plan (DQSP) in the Spring of 1998. An Agency-wide team was assembled and the DQSP was developed and submitted to the Administrator in December 1998. The Strategic Plan recommended both error prevention and error correction strategies. Implementation of an effective error correction process is vital to maintaining the Agency's credibility with its stakeholders.

This document describes standard operating procedures (SOPs) for an integrated error correction process (IECP). The overall framework for this process is depicted in Figure 1.



As laid out in the DQSP, the error correction process should help stakeholders funnel questions and discrepancies into the Agency's information infrastructure, enabling discrepancies to be input through diverse public interfaces and routed to the appropriate data system to begin their trip to the appropriate place and level for resolution. Such a process is not intended to substitute system specific data correction approaches undertaken by the Program Offices. The following describes the standard operating procedures of the IECF.

Step 1. Reporting Discrepancies

Various EPA Web sites will announce that the Agency has developed an IECF, including the Envirofacts Warehouse and other Web sites that display EPA data (e.g. the Sector Facility Indexing Project and various program offices sites). External Web sites will be asked to help publicize the IECF (e.g. ECOS, RTK-NET, and EDF). These portals will direct interested individuals to the IECF Web interface, where details will be provided. The IECF Web interface will reside on Envirofacts, where the majority of program system data disseminated to the public are made available. A "smart" error correction

Discrepancy Reporting Process for Envirofacts

Mr. Johnson works as an environmental compliance specialist for a large wastewater treatment plant. Because Mr. Johnson is concerned about public perceptions regarding his facility, he visits the Envirofacts Warehouse to review data pertaining to his facility. During his review, Mr. Johnson identifies a compliance violation which he believes is incorrect. After checking his own records, he finds documentation indicating that the violation in question was disputed and later overturned. At this point, Mr. Johnson decides to notify EPA regarding this discrepancy. To do this, Mr. Johnson presses the error correction button found at the bottom of the page where he identified the discrepancy. A notification for submitting error corrections appears on his screen. The notification automatically captures key information from the location where the discrepancy was identified within Envirofacts, including the Uniform Resource Locator (URL) and the information associated with his facility. Other key information that the notification automatically captures includes the date and time of the submission. A registration or case number is also assigned automatically. The notification prompts Mr. Johnson for his E-mail address, name, and phone number -- information EPA uses to confirm receipt of his discrepancy notification. The notification then prompts Mr. Johnson to identify the data element in question by clicking on it. The element is automatically copied to the discrepancy field on the notification, and Mr. Johnson confirms that this is the data element of concern. After identifying the appropriate source data system, he provides a brief explanation as to why he feels the information should be changed, and his sense of what the correct value is. Mr. Johnson notes that he may later be called upon to provide certified documentation to substantiate his claim. Mr. Johnson completes the notification in five minutes, after which he presses a button at the bottom of the notification to submit the discrepancy to EPA for resolution. A screen pops up indicating successful transmission to the Agency. Within X business days, Mr. Johnson receives an E-mail from EPA acknowledging his contribution. The message indicates to whom the discrepancy has been forwarded, how the resolution process will

notification will be used to submit discrepancies for resolution. Program offices and hotline operators fielding discrepancies will direct callers to the IECP via the Web when appropriate¹, or provide direct assistance for those without Internet access.

Individuals seeking to resolve discrepancies in data available at other EPA Web sites will use a similar² error correction notification available at the IECP Web interface to provide EPA with the information necessary to take appropriate action. Those without Internet access may call Agency Program Office or Hotline staff who will take down their information using the same notification. Callers who have Internet access will be directed to use the IECP Web interface to submit their discrepancies. In all cases, notifications will be submitted electronically to a Data Quality Action Officer (DQAO). The DQAO is a designee within EPA Headquarters responsible for oversight and management of the error correction process. The system will automatically E-mail a brief message indicating successful transmission of discrepancies.

Step 2. Managing and Routing Discrepancies

All discrepancy notices received will be stored and managed by an error correction tracking tool. The DQAO will screen submissions for their admissibility to the IECP using the following criteria:

- Submission is understandable (if not, request clarification);
- Submission is an error (if question, comment, etc., redirect to appropriate party); and
- Certification information required (if yes, request additional information).

After screening the submission, the DQAO will either:

- A. Redirect it with a referral message to the submitter; or
- B. Forward it to the appropriate party for resolution.

A. Redirect with Referral: In those cases where submissions are not admissible, the DQAO will send contributors a notice indicating this. It is anticipated that some submissions will not be errors, but rather concerns, requests, or questions which can better be addressed by other organizations within the Agency. In those cases, the DQAO will also forward the submission to Program or Regional staff who can address the issue.

B. Forward for Resolution: For admissible discrepancies, the DQAO will review the Error Correction Network Diagram for the system in question, and select the appropriate system action officer (SAO). The DQAO will route admissible discrepancy notifications to SAOs who have program or state-level authorization to make determinations and institute corrections. The SAO will confirm that the discrepancy originates within his/her system. The DQAO will also send an E-mail to submitters (or phone call for those without Internet access) indicating where the discrepancy was routed, how the resolution process will proceed, and the time frame within which resolution can be expected. The DQAO will employ a tracking tool to coordinate all correspondence pertaining to error correction, which will be necessary for execution and effective oversight of the IECP.

The DQAO will have **X** business days from the time of receipt to screen and redirect or refer the submission.

Step 3. Resolving Discrepancies

SAOs will employ procedures currently being used to address discrepancy notifications received (mostly) from the regulated community³. Upon receipt of a notification, SAOs may resolve discrepancies one of two ways:

- A. The SAO may determine to reject changes to data, and notify the submitter of the rejection with appropriate explanation. A copy of the notice will be sent to the DQAO.
- B. For all submissions that result in corrections, the SAO will ensure that changes are affected in the appropriate data system. The SAO will notify the DQAO regarding institution of a flag in Envirofacts⁴, as well as when the change has been instituted in the source data system.

The SAO will have **X** business days from the time of receipt to review the discrepancy, make a determination, and notify the submitter and DQAO.

Step 4. Resolving Flags in Envirofacts and Periodic Review of the IECP

The DQAO will be responsible for oversight of flags in Envirofacts. This includes working with Envirofacts staff to ensure that flags are instituted and removed when appropriate. The DQAO will also ensure that expected changes appear in subsequent data refreshes in Envirofacts. In those instances where this has not occurred, the DQAO will notify the responsible SAO in order to rectify the situation.

The Office of Environmental Information (OEI) will analyze the discrepancy notifications received over a period of time in order to characterize and assess the types of discrepancies that are submitted, systems to which discrepancies apply, and other managerial checks. This information will also be checked against the performance standards developed to measure the systems's functionality and the timeliness of the actions taken. This could be used to provide input for future policies for increasing the quality of the Agency's information.

Management summary reports will be developed on a monthly, quarterly and annual basis.

Endnotes

1. The IECP will address EPA program system data made available to the public. Program office and hotline personnel will use guidance to direct callers elsewhere in those cases where issues raised fall outside the IECP purview. For example, discrepancies associated with aggregated or otherwise manipulated data found in Agency information products will be referred to appropriate program staff where existing approaches for addressing such discrepancies will continue to be employed. The IECP will not replace traditional information collection efforts that capture revisions (e.g. usage of Form R for TRI data).
 2. The notification will be similar to the “smart” error correction notification, but will lack the “screen capture” component of that notification. The screen capture feature could be used once various EPA/external Web sites displaying Agency data are upgraded to allow for the “smart” notification to appear at the location on their Web sites where data are displayed.
 3. As part of the Error Correction Network Diagram, the Agency will need to review current error correction procedures, map them, and develop data revision and concordance guidelines that clarify what to do and where to find individuals with the responsibility and authority to resolve discrepancies. Such guidelines will also establish what data are admissible to the IECP, and rules to follow when considering changes to Agency data.
- The actual process that individual SAOs will use to make determinations regarding whether to accept or reject suggested changes to data will vary by system. Error correction processes used by the Agency will need to become better understood during the institution of the IECP. In some cases, EPA may wish to modify the error correction procedures for some systems to expedite discrepancy resolution.
4. Flags within Envirofacts will be displayed at the “Facility Detail Report” level.